"PROCESSING METHOD AND DEVICE USING SCENE CHANGE DETECTION"

FIELD OF THE INVENTION

The invention relates to a method allowing to automatically detect gradual scene transitions in H.264/AVC video streams. The method is based on the usage of novel coding parameters introduced by H.264, enabling very efficient and cost-effective detection.

BACKGROUND OF THE INVENTION

During the recent years, international video coding standards have played a key role in facilitating the adoption of digital video in various professional and consumer applications. Most influential standards have been developed by two organizations: ITU-T and ISO/IEC MPEG, sometimes jointly (for example : MPEG-2/H.262). The newest joint standard is H.264/AVC, which was expected to be officially approved in 2003 by ITU-T as Recommendation H.264/AVC and by ISO/IEC as International Standard 14496-10 (MPEG-4 Part 10) Advanced Video Coding (AVC). The main goals of the H.264/AVC standardization have been to achieve a significant gain in compression performance and to provide a "network-friendly" video representation addressing "conversational" (telephony) and "non-conversational" (storage, broadcast, streaming) applications. Currently, H.264/AVC is broadly recognized for offering significantly improved rate-distortion efficiency relative to the existing standards, and H.264/AVC-based solutions are also being considered in other standardization bodies, such as the DVB- and DVD-Forum. Implementations of H.264/AVC encoder/decoder are already becoming available, as seeen for instance in : "Emerging H.264 standard : Overview and TMS320C64xDigital Media Platform Implementation – white paper, at : http:///www.ubvideo.com/public. There is also, on the Internet, a growing number of sites offering information about H.264/AVC, among which an official database of ITU-T/MPEG JVT [Joint Video Team] (Official H.264 documents and software of the JVT at : ftp://ftp.imtc-files.org/jvt-experts/) provides free access to documents reflecting the development and status of H.264/AVC, including the draft updates.

The H.264/AVC syntax and coding tools may be recalled here. First, H.264/AVC employs the same principles of block-based motion-compensated transform coding that are known from the established standards such as MPEG-2. The H.264 syntax is, therefore, organized as the usual hierarchy of headers (such as picture-, slice- and macro-block headers) and data (such as motion-vectors, block-transform coefficients, quantizer scale, etc). While

most of the known concepts related to data structuring (e.g. I, P, or B pictures, intra- and inter macro-blocks) are maintained, some new concepts are also introduced at both the header- and the data level. Mainly H.264/AVC separates the Video Coding Layer (VCL), which is defined to efficiently represent the content of the video data, and the Network Abstraction Layer (NAL), which formats data and provides header information in a manner appropriate for conveyance by the higher level (transport) system.

One of the main particularities of H.264/AVC at the data level is also the use of more elaborate partitioning and manipulation of 16x16 macroblocks (a macroblock MB includes both a 16x16 block of luminance and the corresponding 8x8 block of chrominance, but many operations, e.g. motion estimation, actually take only the luminance and project the results on the chrominance). So, the motion compensation process can form segmentations of a MB as small as 4x4 in size, using motion vector accuracy of up to one-fourth of a sample grid. Also, the selection process for motion compensated prediction of a sample block can involve a number of stored previously-decoded pictures, instead of only the adjoining ones. Even with intra coding, it is now possible to form a prediction of a block using previously-decoded samples from neighboring blocks (the rules for this spatial-based prediction are described by the so-called intra prediction modes). After either motion compensated- or spatial-based prediction, the resulting prediction error is normally transformed and quantized based on 4x4 block size, instead of the traditional 8x8 size. This aspect is especially relevant for the invention defined in the following description, and will be highlighted later in the description. The H.264/AVC still uses other specific realizations (e.g. entropy coding), most of which are fixed or can only be altered at or above the picture level.

Concerning the motion compensation, general concepts and particularities of H.264/AVC have also to be recalled. Most established video coding standards, such as MPEG-2, inherently use block-based motion compensation as a practical method of exploiting correlation between subsequent pictures in video. This method attempts to predict each macroblock in a given picture by its "best match" in an adjacent, previously decoded, reference picture. If the pixel-wise difference between a macroblock and its prediction is small enough, this difference (or residue) is encoded rather that the macroblock itself. The relative displacement of the prediction block with respect to the grid position of the actual MB is indicated by a motion vector, which is coded separately. Fig.1 illustrates this for the case of bi-directional prediction, where two reference pictures P_i and P_{i+1} are used, one in the past and one in the future (in the display order). Pictures (such as B_i in Fig.1) that are

predicted in this way are called B-pictures. Otherwise, pictures that are predicted by referring only to the past are called P-pictures.

With H.264/AVC, these basic concepts are further elaborated. Firstly, motion compensation in H.264/AVC is based on multiple reference pictures prediction : a match for a given block can be sought in more distant past or future pictures, instead of only in the adjacent ones. Secondly, H.264/AVC allows to divide a MB into smaller blocks, and to predict each of these blocks separately. This means that the prediction for a given MB can in principle be composed of different blocks, retrieved with different motion vectors and from different reference pictures. The number, size and orientation of the prediction blocks are uniquely determined by the choice of an inter mode. Several such modes are specified, allowing block sizes 16x8, 8x8, etc., down to 4x4.

Another innovation in H.264/AVC allows the motion compensated prediction signal to be weighted and offset by amounts specified by the encoder. This means that in the case of a bi-directional prediction concerning a frame B(i) predicted from previous frames P(i-n) and P(i-1) and following frames P(i+j) and P(i+m), the encoder can choose unequal amounts by which the prediction blocks from the past and the prediction blocks from the future will contribute in the total prediction. This feature allows to dramatically improve coding efficiency for scenes containing fades.

The problem is however the following one. Recent advances in computing, communications and digital data storage have led to a tremendous growth of large digital archives, characterized by a steadily increasing capacity and content variety. Finding efficient ways to quickly retrieve stored information of interest is therefore of crucial importance. Since searching manually through terabytes of unorganized stored data is tedious and time-consuming, there is a growing need to transferring information search and retrieval tasks to automated systems. Search and retrieval in large archives of unstructured video content are usually performed after said content has been indexed using content analysis techniques. These techniques are based on algorithms such as image processing, pattern recognition and artificial intelligence, which aim at automatically creating, in view of the description of said video content, annotations of video material (such annotations vary from low-level signal related properties, such as color and texture, to higher level information, such as presence and location of faces).

One the most important content descriptors is the shot boundary indicator, as seen for instance in a document such as the international patent application WO 01/03429 (PHF99593). A shot is a video segment that has been taken using continuously a single

camera, and shots are generally considered as the elementary units constituting a video. Detecting shot boundaries thus means recovering those elementary video units, which in turn provide the ground for nearly all existing video abstraction and high-level video segmentation algorithms (see for instance the document "Video abstracting", by R.Lienhart and al, Communications of the ACM, 40(12), 1997, pp.55-62).

During video editing, shots are connected using shot transitions, that can be classified into at least two classes : abrupt transitions and gradual transitions. Abrupt transitions, also called hard cuts and obtained without any modifications of the two shots, are fairly easy to detect, and they constitute the majority in all kind of video productions. Gradual transition, such as fades, dissolves and wipes, are obtained by applying some transformation to the two involved shots. During video production, each transition type is chosen carefully in order to support the content and context of the video sequences. Automatically recovering all their positions and types, therefore, may help a machine to deduce high-level semantics. For instance, in feature films, dissolves are often used to convey a passage of time. Also dissolves occur much more often in feature films, documentaries, biographical and scenic video material than in newscasts, sports, comedy and shows. The opposite is true for wipes. Therefore, the automatic detection of transitions and their type can be used for automatic recognition of video genre.

Because of the large application area for the upcoming H.264/AVC standard, there will be a growing demand for efficient solutions for H.264/AVC video content analysis. During the recent years, several efficient content analysis algorithms and methods have been demonstrated for MPEG-2 video, that almost exclusively operate in the compressed domain. Most of these methods could be extended to H.264/AVC, since H.264/AVC in a way specifies a superset of MPEG-2 syntax, as indicated above. However, due to the limitations of MPEG-2, these existing methods may not give adequate or reliable performance, which is a deficiency that is typically addressed by including additional and often costly methods operating in the pixel- or audio domain.

SUMMARY OF THE INVENTION

It is therefore an object of the invention to propose a method allowing to avoid said drawback in all the situations where weighted predictions of frames take place with an unequal amount of prediction from the past and the future of the frame to be predicted.

To this end, the invention relates to a method of processing digital coded video data available in the form of a video stream consisting of consecutive frames divided into

macroblocks, said frames including at least I-frames, independently coded, P-frames, temporally disposed between said I-frames and predicted from at least a previous I- or P-frame, and B-frames, temporally disposed between an I-frame and a P-frame, or between two P-frames, and bidirectionally predicted from at least these two frames between which they are disposed, said predictions being performed by means of a weighted prediction with unequal amount of prediction from the past and the future, said processing method comprising the steps of :

- determining for each successive macroblock of the current frame related coding parameters characterizing, if any, said weighted prediction ;
- collecting said parameters for all the successive macroblocks of the current frame, for delivering statistics related to said parameters ;
- analyzing said statistics for determining a change of preference for the direction of prediction ;
- detecting the occurrence of a gradual scene change in the sequence of frames each time a change of preference has been determined.

More precisely, according to the invention, the analysis step is provided for comparing the number of macroblocks having the same directional preference and similar weighting against a predefined threshold derived in relation to the total number of macroblocks in the current frame. Preferably, an information about the location and the duration of each scene change is produced and stored in a file.

Another object of the invention is to propose a processing device allowing to carry out the method defined above.

To this end, the invention relates to a device for processing digital coded video data available in the form of a video stream consisting of consecutive frames divided into macroblocks, said frames including at least I-frames, independently coded, P-frames, temporally disposed between said I-frames and predicted from at least a previous I- or P-frame, and B-frames, temporally disposed between an I-frame and a P-frame, or between two P-frames, and bidirectionally predicted from at least these two frames between which they are disposed, said predictions being performed by means of a weighted prediction with unequal amount of prediction from the past and the future, said device comprising the following means :

- determining means, provided for determining for each successive macroblock of the current frame related coding parameters characterizing, if any, said weighted prediction ;

- collecting means, provided for collecting said parameters for all the successive macroblocks of the current frame, for delivering statistics related to said parameters ;
- analyzing means, provided for analyzing said statistics for determining a change of preference for the direction of prediction ;
- detecting means, provided for detecting the occurrence of a gradual scene change in the sequence of frames each time a change of preference has been determined.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will now be described, by way of example, with reference to the accompanying drawings in which :

- Fig.1 illustrates a conventional example of bidirectional prediction ;
- Fig.2 illustrates the basic principle of the weighting prediction for a B-frame, in the case of the H.264/AVC standard ;
- Fig.3 is a block diagram of an implementation of the processing method according to the invention.

DETAILED DESCRIPTION OF THE INVENTION

As explained above when recalling the general concepts and particularities of H.264/AVC concerning the motion prediction, the motion-compensated prediction signal can be weighted by amounts specified by the encoder. A weighted prediction can be used to achieve bi-directional prediction (B-pictures) where the prediction block from the past and that from the future are present in the total prediction by unequal amounts (with MPEG-2, this is limited to one possibility of weighting both prediction signals by a factor of ½).

The principle of the invention is that, because of this inequality, the presence of a gradual shot transition can be indicated by a gradual change in the preference for prediction from one direction to the other. Such a change of preference for the direction of prediction can be detected by analyzing the statistics of related coding parameters characterizing weighted prediction. For example, this analysis can include comparing the number of macroblocks having the same directional preference and similar weighting against a given threshold, which could be derived in relation to the total number of macroblocks in the picture. Furthermore, (local) uniformity of distribution of such macroblocks can be examined to make sure that the change in directional preference for prediction is indeed a consequence of a gradual scene transition. Also, some additional analysis may be performed to take into

account the possible use of sub-macroblock motion prediction, and in weighted prediction, as is allowed e.g. in H.264/AVC.

An example of bidirectional prediction in e.g. H.264/AVC is illustrated in Fig.2, showing the prediction of a picture B_i from previous and following pictures P_{i-n}, P_{i-1}, P_{i+j}, P_{i+m}. The prediction for a macroblock MB, called MB_{Pred} and equal to B_1 "+" B_2 "+" B_3, with $B_1 = alpha1.b_1 + alpha2.b_2$ (where alpha 1 and alpha 2 are coefficients), is composed of three prediction blocks, such that the lower half of the macroblock MB_{Pred} is predicted by two 8 x 8 blocks B_2 and B_3, and the upper half by one 8 x 16 blocks B_1. Each of these prediction blocks pertains to a different reference picture and has a distinct motion vector MV, as allowed in H.264. Unlike B_2 and B_3, the block B_1 is obtained using weighted prediction, i.e. it is obtained by performing the sum of two blocks b_1 and b_2 that are present in the sum by unequal amounts controlled by corresponding weighting parameters alpha1 et alpha2. The statistics of these weighting parameters (absolute value and sign) are collected for all macroblocks, and the statistics distribution over the plurality of macroblocks is analyzed to achieve the detection of gradual scene transitions.

An implementation of the processing method according to the invention is shown in the block diagram of Fig.3, that illustrates for example in the case of an H.264/AVC bitstream the concept previously explained, said example being however not a limitation of the scope of the invention. In the illustrated decoding device, a demultiplexer 21 receives a transport stream TS and generates demultiplexed audio and video streams AS and VS. The video stream is received by an H.264/AVC decoder 22, for delivering as usually a decoded video stream DVS. Said decoder 22 mainly comprises an inverse quantization circuit 221 (Q^{-1}), an inverse transform circuit 222 (T^{-1}), which is in the present case an inverse DCT circuit, and a motion compensation circuit 223. It also comprises a so-called Network Abstraction Layer Unit (NALU) 224 , provided for collecting the received coding parameters that characterize the weighted predictions performed (for instance, some relevant coding parameters may be "luma_weight", "luma_offset", "luma_log2_weight_denom", etc, which are used in equations that characterize weighting and offset of the prediction samples). The output signals of said unit 224 are weighted prediction parameter statistics WPPS that are received by an analysis circuit 23 for suitable processing. The processing operation carried out in the circuit 23 then produces an information about location and duration of gradual scene changes in the stream originally received, and this information is then stored in a file 24, e.g. in the form of the commonly used CPI (Characteristic Point Information) table. This

output information is now available for applications such as video summarization, automatic chaptering, etc.

It can be added that there are numerous ways of implementing functions by means of items of hardware or software (the method of the invention can then be carried out by a computer program product for a processing unit comprising a set of instructions which, when loaded into said processing unit, causes this processing unit to carry out the method as described above), or both. In this respect, the drawings are very diagrammatic and represent only one possible embodiment of the invention. Thus, although a drawing (in the present case, Fig.3) shows different functions as different blocks, this by no means excludes that a single item of hardware or software carries out several functions. Nor does it exclude that an assembly of items of hardware or software or both carry out a function.. These remarks are intended to recall that the detailed description, with reference to the drawings, illustrates rather than limits the invention and that there are numerous alternatives, which fall within the scope of the appended claims. The word "comprising" does not exclude the presence of other elements or steps that those listed in a claim. The word "a" or "an" preceding an element or step does not exclude the presence of a plurality of such elements or steps.